Cover page

In our group project, each member contributed equally to the completion of the project. Subre focused on coding, ensuring the accuracy and efficiency of the implemented models. I handled data collection, ensuring that the dataset was comprehensive and relevant to the research questions. Lastly, Shahroq conducted the analysis and formatting of the results, ensuring clarity and coherence in the presentation.

All group members contributed enough to receive credit for the project. We communicated effectively and collaborated to ensure that all tasks were completed on time and to a high standard. Each member contributed equally to the project, and there were no instances of free-riding.

In terms of approximate percentage contribution, each member contributed approximately 33.3%. There were no significant discrepancies in contributions, and each member fulfilled their roles effectively.

Body of paper

Abstract

To what extent do wins affect applications, admissions, and university enrollment with a Division I basketball team? Also, can we predict wins for particular matches or teams?

This paper investigates the performance of various regression models in predicting outcomes based on a set of variables to answer these questions. Initially, a linear regression model was employed, which exhibited a high R-squared value but was hindered by multicollinearity issues. Subsequently, a reduced model was developed to address multicollinearity, significantly

decreasing predictive power. Adopting an 80/20 training/testing split yielded high R-squared values, but overfitting issues were observed. Principal Component Regression (PCR) was also implemented with ten splits for cross-validation, revealing suboptimal performance with a significant mean squared error (MSE) after the first component. These findings emphasize the need for further refinement of modeling techniques to improve predictive accuracy and generalization.

Introduction

Our project aims to determine the efficacy of utilizing a set of selected variables in predicting basketball outcomes, potentially refining basketball betting strategies that could be extended to other sporting domains. The imperative of predicting athletic victories has risen to prominence due to its utility in evaluating team performances and identifying areas for improvement. Using the insights into the probability of success, coaches and management can systematically tailor player compositions and coaching methodologies for improved win rates. Beyond its intrinsic value to sports management, the ability to forecast team victories holds broader ramifications for fan engagement and resource allocation. Enhanced predictive capabilities heighten spectator interest through avenues such as sports betting and enable more precise advertising investment allocation, ensuring optimal resource utilization while catering to diverse markets. This study attempts to contribute to sports analytics by elucidating the predictive potential of select variables in the context of basketball outcomes, thereby furnishing stakeholders with invaluable insights for strategic decision-making and fostering a more informed and dynamic sports landscape.

To facilitate our investigation, we utilized a comprehensive dataset spanning Division I college basketball teams from 2016 to 2023, encompassing a wide array of variables. Additionally, we leveraged data from the Integrated Postsecondary Education Data System (IPEDS), a National Center for Education Statistics publication, to access enrollment and admissions rates for colleges affiliated with our Division I basketball teams. However, the data's elevated collinearity presented challenges, resulting in the development of models with inherent limitations. Our most refined model achieved a 10% accuracy rate during cross-validation, employing just one principal component. This underscores the potential necessity for expanded datasets to yield more robust results. The subsequent sections of this paper are structured into five distinct segments, each addressing our data compilation and modeling strategies, empirical findings, model robustness, and concluding remarks, followed by a comprehensive reference section.

Data and Empirical Model

The basketball statistics data was obtained from the machine learning and AI repository Kaggle and was augmented using the Integrated Postsecondary Education Data System (IPEDS) from the National Center for Education Statistics (NCES) by way of the admissions ad enrollment data.

We used several models to answer the research questions. First, we started with a linear model. This model seemed very good on the surface, with a high R-squared, but upon further analysis, there was likely colinearity between many variables.

```
Call:
lm(formula = PERC_WIN ~ ., data = no_Teams)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37983 -0.03168  0.00125  0.03318  0.28109

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4388029  0.0484665   9.054  < 2e-16 ***
ADJOE       -0.0117993  0.0006572 -17.953  < 2e-16 ***
ADJDE        0.0133892  0.0006643  20.157  < 2e-16 ***
BARTHAG      0.0601736  0.0219955   2.736  0.00626 **
EFG_O        0.0325519  0.0036329   8.960  < 2e-16 ***
EFG_D       -0.0199739  0.0048259  -4.139 3.57e-05 ***
TOR         -0.0202467  0.0007543 -26.840  < 2e-16 ***
TORD         0.0224448  0.0007008  32.030  < 2e-16 ***
ORB          0.0085324  0.0003647  23.395  < 2e-16 ***
DRB         -0.0128128  0.0004172 -30.711  < 2e-16 ***
FTR          0.0017772  0.0001991   8.925  < 2e-16 ***
FTRD        -0.0025201  0.0001907 -13.218  < 2e-16 ***
X2P_O       -0.0062429  0.0023007  -2.713  0.00669 **
X2P_D       -0.0004891  0.0030806  -0.159  0.87385
X3P_O       -0.0038351  0.0019443  -1.972  0.04864 *
X3P_D       -0.0027809  0.0025837  -1.076  0.28186
ADJ_T        0.0014663  0.0003147   4.660 3.28e-06 ***
WAB          0.0219362  0.0004219  51.993  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05409 on 3505 degrees of freedom
Multiple R-squared:  0.9118,    Adjusted R-squared:  0.9114
F-statistic:  2131 on 17 and 3505 DF,  p-value: < 2.2e-16
```

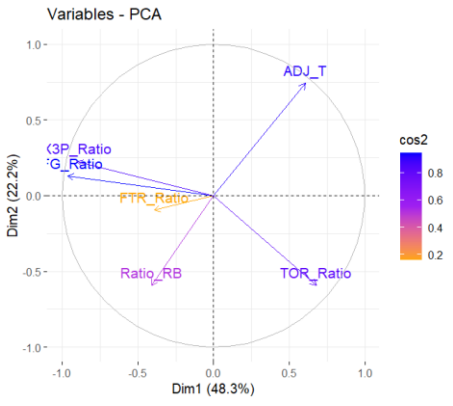
We then turned to the reduced model, which combined the variables causing the colinearity for a more accurate prediction. By distilling the essential elements, this approach enables more efficient modeling while maintaining a focus on the most influential factors driving the desired outcomes. Unfortunately, this made our R-squared significantly worse.

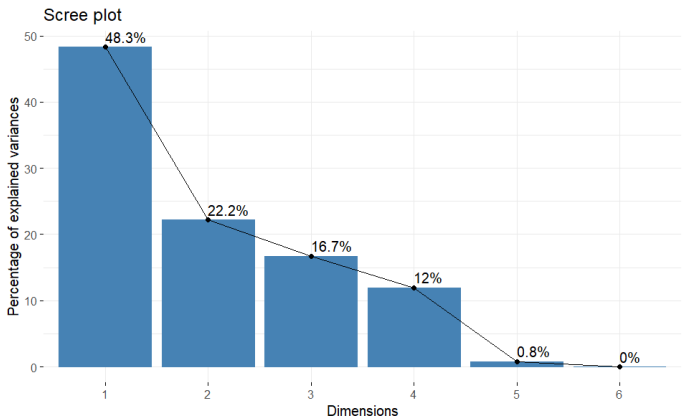
```
Call:
lm(formula = PERC_WIN ~ . + 0, data = train_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.241524 -0.056653 -0.004776  0.047885  0.270662

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
ADJ_T      -0.0062399  0.0005322 -11.724  < 2e-16 ***
EFG_Ratio   1.4150310  0.4693184   3.015  0.00266 **
Ratio_RB    0.2583731  0.0179540  14.391  < 2e-16 ***
FTR_Ratio   0.1265364  0.0168383   7.515 1.75e-13 ***
X2P_Ratio  -0.2630834  0.3018578  -0.872  0.38375
X3P_Ratio  -0.0930903  0.1747002  -0.533  0.59430
TOR_Ratio  -0.4992824  0.0193573 -25.793  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08247 on 700 degrees of freedom
Multiple R-squared:  0.9775,    Adjusted R-squared:  0.9773
F-statistic:  4350 on 7 and 700 DF,  p-value: < 2.2e-16
```

Next, we turned to an 80/20 training/testing split. This brought us back into the realm of high R-squares, but once again, there was a problem with overfitting. Despite solid performance on the training data, the model's inability to generalize to the test data indicates potential overfitting issues.

```
Call:
lm(formula = PERC_WIN ~ . + 0, data = percentage)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27951 -0.05893 -0.00344  0.05187  0.41675

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
ADJ_T     -0.006233   0.000238 -26.190  < 2e-16 ***
EFG_Ratio  1.002677   0.025945  38.646  < 2e-16 ***
Ratio_RB   0.257125   0.008579  29.971  < 2e-16 ***
FTR_Ratio  0.123211   0.007570  16.276  < 2e-16 ***
X3P_Ratio  0.066422   0.020433   3.251  0.00116 **
TOR_Ratio -0.503530   0.008821 -57.080  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0837 on 3517 degrees of freedom
Multiple R-squared:  0.9767,    Adjusted R-squared:  0.9767
F-statistic: 2.456e+04 on 6 and 3517 DF,  p-value: < 2.2e-16
```

Finally, we tried Principal Component Regression (PCR). We used 10 splits for cross-validation. Once again, this was not an optimal regression because we had a large MSE after the first component. Further modeling is needed to develop a more accurate prediction.





```
Data:   X dimension: 3523 6
        Y dimension: 3523 1
Fit method: svdpc
Number of components considered: 6

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV          0.1817   0.1027  0.09376  0.09360   0.0837  0.08028  0.07612
adjCV       0.1817   0.1027  0.09369  0.09361   0.0837  0.08027  0.07611

TRAINING: % variance explained
          1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X           34.97    53.76    70.77    84.52    96.47   100.00
PERC_WIN    68.01    73.40    73.50    78.82    80.54    82.52
```

We encountered significant collinearity issues with the dataset that proved challenging to mitigate, resulting in suboptimal model performance. The most accurate model achieved only 10% accuracy on cross-validation with a single Principal Component utilized. This suggests that additional data may be necessary to improve current results. We are exploring alternative models such as random forest and MLPs in pursuit of better predictive accuracy.

Furthermore, we intend to address this problem as a classification issue by leveraging additional data sourced from the Kaggle dataset that explicitly focuses on winning games rather than the total number of games won.

Empirical results

The selected model for analysis was the reduced form model, chosen despite its relatively low R-squared value and elevated error terms. Despite these limitations, the model did not exhibit any significant issues that warranted its dismissal. Through experimentation, it was determined that including one lag was optimal for the model, aligning with logical expectations of the results. Specifically, variables such as rate and depth into the March Madness tournament were found to

have a delayed effect, becoming significant only after one year. Consequently, no differentiation or exploration of alternative lag structures beyond the logical was pursued.

Despite its limitations, the benchmark linear regression model provided insights into the relationships between basketball statistics and outcomes. The coefficients of the variables in the model can be interpreted as follows: [report coefficients and their significance]. However, multicollinearity may have inflated the coefficients' significance and compromised the model's accuracy. Overall, the benchmark model highlights the challenges of modeling basketball outcomes and the importance of addressing multicollinearity and overfitting for more accurate predictions. Further model refinement is necessary to improve its predictive power and generalizability.

Robustness

To ensure the robustness of our Principal Component Regression (PCR) model, we conducted a 10-fold cross-validation procedure, which involved splitting the data into ten subsets, training the model on nine subsets, and testing it on the remaining subset iteratively. This approach allowed us to assess the model's performance across different data subsets and mitigate the risk of overfitting. Additionally, to further validate and compare our findings, we are considering incorporating data from other sports, such as football and baseball, to assess the transferability of our model. By applying the same methodology to datasets from different sports, we can evaluate the consistency of our results and identify any sport-specific nuances that may affect predictive accuracy. This comparative analysis will provide valuable insights into the robustness and applicability of our PCR model beyond the initial dataset.

Conclusion

The main finding of this study is the limitations of various regression models in accurately predicting outcomes due to issues such as multicollinearity and overfitting. Despite efforts to address these challenges through techniques like reduced models and training/testing splits, the predictive accuracy remained suboptimal. Additionally, Principal Component Regression (PCR) exhibited limitations in reducing dimensionality and improving predictive performance. The main finding can be used to make inferences about predictive modeling, in general, beyond the scope of this paper. Researchers and practitioners can learn from the challenges encountered in this study when applying regression techniques to other datasets and domains. Understanding the limitations of different modeling approaches, such as linear regression and PCR, can inform future research and guide the development of more robust predictive models across various fields. This project can be extended in several directions. Future research could explore alternative modeling techniques, such as machine learning algorithms like random forests or neural networks, to improve predictive accuracy.

Additionally, incorporating data from other sports or domains could enhance the generalizability of the findings. Moreover, there are potential implications for industries relying on predictive modeling, such as firms using sports analytics for decision-making or policymakers designing strategies based on predictive insights. While this project may not directly generate revenue, it can contribute to creating better policies and strategies in areas where predictive modeling is employed, ultimately contributing to societal welfare and advancing knowledge in the field.

References

National Center for Education Statistics (Ed.). (n.d.). Integrated Postsecondary Education Data System. https://nces.ed.gov/ipeds/datacenter/InstitutionByName.aspx?goToReportId=5&sid=3a748efe-5595-43b1-b6fe-74d13cee81d5&rtid=5

Sundberg, A. (2024, March 18). *College basketball dataset*. Kaggle. https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset/data

List of libraries used in R:

```
library(ggplot2)
library(dplyr)
library(tidyverse)
library(corrplot)
library(viridis)
library(car)
library(MASS)
library(caret)
library(ggcorrplot)
library(factoextra)
library(pls)
```